American Journal of
EPIDEMIOLOGY

Volume 154
Number 8
October 15, 2001

Sponsored by the Society for Epidemiologic Research
Published by Oxford University Press

ORIGINAL CONTRIBUTIONS

Limitations of the Case-only Design for Identifying Gene-Environment Interactions

Paul S. Albert,[1] Duminda Ratnasinghe,[2] Joseph Tangrea,[2] and Sholom Wacholder[3]

The case-only design, which requires only diseased subjects, allows for estimation of multiplicative interactions between factors known to be independent in the study population. The design is being used as an alternative to the case-control design to study gene-environment interactions. Estimates of gene-environment interactions have been shown to be very efficient relative to estimates obtained with a case-control study under the assumption of independence between the genetic and environmental factors. In this paper, the authors explore the robustness of this procedure to uncertainty about the independence assumption. By using simulations, they demonstrate that inferences about the multiplicative interaction with the case-only design can be highly distorted when there is departure from the independence assumption. They illustrate their results with a recent study of gene-environment interactions and risk of lung cancer incidence in a cohort of miners from the Yunnan Tin Corporation in southern China. Investigators should be aware that the increased efficiency of the case-only design is a consequence of a strong assumption and that this design can perform poorly if the assumption is violated. *Am J Epidemiol* 2001;154:687–93.

case-control studies; cohort studies; environment; genes; genetics

Piegorsch et al. (1) showed that under the assumption of independence between two factors in the population, efficient estimates of interaction (departure from multiplicative risk ratios) can be obtained without studying any controls. Several authors have proposed the use of case-only designs as an alternative to case-control designs to study gene-environment interactions (1–6), and the case-only design has been used in a number of studies (7–10). Although the assumption of independence for a gene and an external environmental factor or behavior seems reasonable, there will rarely be sufficient empirical data to support the independence assumption. This problem led us to examine the robustness of inferences to violations of independence.

The specific motivation for this work was a recent study examining the association between polymorphisms of the DNA repair gene *XRCC1* and lung cancer risk in a cohort of high-risk tin miners from the Yunnan Tin Corporation (YTC) in southern China. By using a nested case-control study design, Ratnasinghe et al. (11) explored the association between polymorphisms of *XRCC1* and the risk of lung cancer and examined whether selected environmental exposures might modify this association.

In this paper, we explore the properties of the case-only design using theoretical arguments and an empirical study of a gene, smoking, and lung cancer. Ratnasinghe et al. (11) originally used a standard case-control analysis to assess gene-environment interactions. We explore the case-only analysis of the same study, which would be available from a case-only design, as an alternative to the case-control design. In addition, we examine the sensitivity of the case-only design to departures from independence.

Four possible strategies for screening for gene-environment interactions are considered:

Received for publication November 9, 2000, and accepted for publication May 16, 2001.

Abbreviations: E, environmental exposure; G, genetic mutation or polymorphism; YTC, Yunnan Tin Corporation.

[1] Biometric Research Branch, Division of Cancer Treatment and Diagnosis, National Cancer Institute, Bethesda, MD.

[2] Cancer Studies Prevention Branch, Division of Clinical Sciences, National Cancer Institute, Bethesda, MD.

[3] Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD.

Reprint requests to Dr. Paul S. Albert, 15158 Callohan Court, Silver Spring, MD 20906 (e-mail: albertp@ctep.nci.nih.gov).

1. Cohort design: The interaction effect is estimated from incident cases and all noncases, as in a prospective cohort.
2. Case-control design: The interaction effect is estimated from incident cases and controls selected from among all noncases, as in a nested case-control study.
3. Case-only design: The interaction effect is estimated from cases only.
4. Case-control/cohort design with an adaptive case-only analysis: The interaction effect is estimated from either the standard case-control analysis or a case-only analysis based on the evidence for gene-environment independence in the controls.

In this paper, we explore these four strategies for estimating multiplicative interaction under various assumed gene-environment associations.

THE CASE-ONLY DESIGN

In this section, we review the case-only design and discuss its efficiency relative to the cohort and case-control designs. We also examine the sensitivity of the case-only design to the independence assumption.

Justification of the case-only design

We assume that the genetic mutation and environmental exposure are binary variables, unless noted otherwise. Let D, G, and E indicate presence of disease, the genetic mutation or polymorphism, and environmental exposure, respectively. We also assume that there are no confounding factors. The multiplicative interaction is defined as

$$I = \frac{\Psi_{GE}}{\Psi_G \Psi_E}, \qquad (1)$$

where Ψ_{GE} is the ratio of the odds of disease for those persons positive for both the genetic ($G+$) and environmental ($E+$) factors relative to the odds for those negative for both factors (G–, E–). Similarly, Ψ_G is the ratio of the odds of disease for $G+$ and E– relative to G– and E–. The parameter Ψ_E is defined analogously for E. For dichotomous G and E, I can be estimated by using equation 1. For categorical or continuous G or E, the multiplicative interaction I is equivalent to β_3 in a standard logistic regression model

$$\text{logit}\, P(D = 1) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE. \qquad (2)$$

Piegorsch et al. (1) showed that under the assumption of independence between the environmental factor and genetic marker and under a rare-disease assumption ($P(D = 1|G = i)$ is negligible for $i = 0$ and 1), the interaction effect I can be estimated without studying any controls. Specifically, they show that

$$I_{CO} = I \times \Psi, \qquad (3)$$

where Ψ is the odds ratio relating the genetic marker and environmental exposure in the controls, and I_{CO} is the odds ratio relating the exposure variable with the genetic marker in the cases. Since $I_{CO} = I$ when $\Psi = 1$, we can estimate the interaction odds ratio with the case-only design.

Table 1 introduces notation for the frequency of cases and controls by dichotomous classifications of gene and exposure. In this situation, the case-only estimate of interaction is $\widehat{I_{CO}} = AX/BW$. A chi-square test of independence can be used to assess the significance of this interaction. Any data from controls in a case-control study or any other source can be used to test for independence by using a chi-square test. Specifically, we can estimate the gene-environment association in controls as $\hat{\Psi} = CZ/DY$.

More generally, when E is categorical or continuous, the case-only estimate of interaction can be obtained by fitting the following logistic regression model to the case data:

$$\text{logit}\, P(G = 1) = \gamma_0 + \gamma_1 E. \qquad (4)$$

Here, $I_{CO} = \exp(\gamma_1)$. Under these conditions, we can also assess the independence assumption by fitting the following logistic regression model to the control data:

$$\text{logit}\, P(G = 1) = \eta_0 + \eta_1 E, \qquad (5)$$

where a test of whether $\eta_1 = 0$ is a test of the independence assumption, since $\psi = \exp(\eta_1)$.

Efficiency of the case-only design

This section demonstrates that asymptotically (with large samples), a gene-environment interaction can be estimated more precisely with a case-only design than with either a cohort or a case-control study. Similar to table 1, let A, B, C, D, W, X, Y, and Z be counts of persons cross-classified by D, E, and G for a cohort study. In a case-control study, C^*, D^*, Y^*, and Z^* will replace C, D, Y, and Z and represent appropriately selected subsets of the noncases in each category defined

TABLE 1. Notation* for the frequency of cases and controls, by dichotomous classifications of gene (G) and exposure (E)

	E+	*E–*
	G–	
Cases	*A*	*B*
Controls	*C*	*D*
	G+	
Cases	*W*	*X*
Controls	*Y*	*Z*

* Defined for a case-control study.

by E and C in the cohort. The asymptotic variance of $\hat{\beta}_3$ from a cohort study is

$$\mathrm{Var}(\hat{\beta}_3) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} + \frac{1}{W} + \frac{1}{X} + \frac{1}{Y} + \frac{1}{Z}. \quad (6)$$

The variance of $\overset{\frown}{\hat{\beta}}_3$, the corresponding estimate from a case-control study, is

$$\mathrm{Var}(\overset{\frown}{\hat{\beta}}_3) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C^*} + \frac{1}{D^*} + \frac{1}{W} + \frac{1}{X} + \frac{1}{Y^*} + \frac{1}{Z^*}. \quad (7)$$

Equations 6 and 7 follow by noting that the asymptotic variance of $\log_e\hat{\psi}$, where $\hat{\psi}$ is the cross-product estimate of the odds ratio in a 2×2 table, can be estimated as the sum of the reciprocals of the cell entries (12) and that estimates of β_3 are differences between independent estimates of this type. The variance of $\hat{\gamma}_1$ from a case-only design is

$$\mathrm{Var}(\hat{\gamma}_1) = \frac{1}{A} + \frac{1}{B} + \frac{1}{W} + \frac{1}{X}. \quad (8)$$

Thus, asymptotically, the case-only design will provide more efficient estimates of gene-environment interaction than either the case-control or the cohort design. For cohort studies of rare diseases, such as the YTC study, the reciprocals of C, D, Y, and Z will be negligible, so $\mathrm{Var}(\hat{\gamma}_1) \approx \mathrm{Var}(\hat{\beta}_3)$ and the efficiency of the case-only and cohort designs will be essentially the same.

A useful way to think of the efficiency gain in the case-only design is that the assumption of independence alleviates the need to estimate the dependence between G and E, thereby eliminating some terms in the variances given in equations 6 and 7. That is, efficiency is gained by relying on the assumption of independence. However, inferences may be highly distorted when the independence assumption is not correct, as illustrated in the discussion that follows.

Sensitivity of the case-only design to the independence assumption

Examination of equation 3 provides some insight into the importance of the independence assumption. Taking logarithms of both sides of equation 3 gives us

$$\log_e I_{CO} = \log_e I + \log_e \psi. \quad (9)$$

Thus, any nonzero value of $\log_e\psi$ (departures from zero provide evidence for a lack of independence) translates to bias for estimating the log-transformed multiplicative interaction ($\log_e I$) when a case-only design is used.

EXAMPLE OF A GENE-ENVIRONMENT INTERACTION STUDY

Screening for gene-environment interactions is important for investigating new mechanisms for disease. For illustration, we use a recent report by Ratnasinghe et al. (11) on the association of polymorphisms of a DNA repair gene and environmental exposure with the incidence of lung cancer in the YTC study; details are provided in the original report. This study was nested in a prospective cohort of 9,143 participants enrolled in 1992, with annual follow-up through 1999. DNA was obtained from 108 cases and 210 controls (6 controls were excluded from the analysis since there was insufficient DNA to assess genotype). The controls were selected from members of the cohort study who were alive and free of cancer at the time the matched case was diagnosed.

We focus on interactions between the single nucleotide polymorphism at codon 194 on the *XRCC1* gene and the level of tobacco use (dichotomized at the median) and alcohol consumption (ever consumption). Approximately 50 percent of study participants were alcohol drinkers, and almost all alcohol consumed was in the form of grain alcohol. Alcohol consumption was dichotomized into ever consumption versus never consumption, since the questionnaire used was designed to accurately reflect drinking status, not actual consumption. These two environmental exposures were chosen from a list of 10 considered by Ratnasinghe et al. (11) to illustrate the most extreme situation of empirical evidence for (alcohol consumption) and against (tobacco use) the independence assumption. The frequency of cases and controls by the *XRCC1* genotype (wild-type vs. variant) and tobacco use and by *XRCC1* genotype and alcohol consumption are given in table 2.

TABLE 2. Frequency of cases and controls, by *XRCC1* genotype and tobacco use and by *XRCC1* genotype and alcohol consumption, in the Yunnan Tin Corporation study, southern China, 1992–1999

	Tobacco use	
	>Median	≤Median
	G–/XRCC1 wild-type genotype	
No. of cases	32	20
No. of controls	29	56
	G+/XRCC1 variant genotype	
No. of cases	33	23
No. of controls	64	61
	Alcohol consumption	
	Yes	No
	G–/XRCC1 wild-type genotype	
No. of cases	31	21
No. of controls	37	48
	G+/XRCC1 variant genotype	
No. of cases	23	33
No. of controls	62	63

Table 3 presents the results (unconditional logistic regression) for gene-environment interactions in which a case-control analysis and a case-only analysis are used. The case-control analyses results demonstrate that the effect of alcohol is significantly less in those persons with the *XRCC1* variant genotype compared with the wild-type genotype ($\hat{I} = 0.37$). Although not statistically significant, there is some indication that the effect of tobacco use may be less in those persons with the *XRCC1* variant genotype compared with the wild-type genotype ($\hat{I} = 0.44$). The corresponding case-only estimates are 0.90 and 0.47, indicating that the inferences for assessing gene-environment interactions may be very different for the case-control and case-only designs. The extreme tobacco use discrepancy is related to the observation among the controls that *XRCC1* variant genotype carriers are apparently heavier smokers than those with the wild-type genotype (the odds ratio assessing independence was estimated as 2.03). Note that equation 3 can be empirically demonstrated by observing that for tobacco use, $0.897 = 0.443 \times 2.03$. The odds ratio assessing independence between alcohol consumption and *XRCC1* was closer to 1 (estimate = 1.28), and inferences about the alcohol-by-gene interaction made with the case-only analysis were similar to those made with the case-control analysis.

Analysis of the YTC study data led us to question the robustness of inferences about the gene-by-environment interactions in the case-only analysis to the lack of independence between the environmental exposure and the genetic marker in the controls.

PERFORMANCE OF THE CASE-ONLY DESIGN

In this section, our focus is on evaluating the operating characteristics of estimation and testing of the multiplicative interaction effect with the case-only design relative to both traditional cohort and case-control designs. Design parameters, as well as parameter estimates from logistic regression models assessing the tobacco-use-by-*XRCC1*-interaction in the YTC cohort, were used to set the following realistic parameter values for our simulations:

- The incidence of lung cancer over follow-up is 3.7 percent (339 lung cancer cases in the cohort of 9,142).
- The prevalence of the *XRCC1* variant genotype is 60 percent.
- The prevalence of the environmental factor is 50 percent.

The following designs are compared:

1. Cohort design: Exposure and genetic testing are performed on all subjects (9,100 subjects) at baseline. The gene-environment β_3 interaction is estimated by fitting the logistic regression model (equation 2) to the full cohort data.
2. Case-control design: Exposure and genetic information is obtained on all cases ($n = 340$) and an equal number of controls. The gene-environment interaction β_3 can be estimated by fitting the logistic regression model (equation 2) to the case-control data.
3. Case-only design: Exposure and genetic information is obtained on all cases ($n = 340$). The interaction effect is estimated by fitting the logistic regression model (equation 4) to the data on cases.
4. Case-control/cohort design with a case-only analysis: Exposure and genetic information is obtained on all cases ($n = 340$) and an equal number of controls. The independence assumption is evaluated by fitting equation 5; if this effect is not significant at the 0.05 significance level, the interaction effect will be estimated by fitting equation 4. Otherwise, a case-control analysis (using equation 2) will be performed to test for the interaction.

   Similar to our example, we also examine this strategy with a 2-1 ratio of controls to cases (340 cases and 680 controls). We also examine the properties of this procedure when we chose a significance level of 0.15 for testing for independence and when the decision to conduct a case-control or case-only analysis was based on the size of the odds ratio for the gene-environment association.

All testing was performed by using two-sided Wald tests (i.e., a test based on $\hat{\beta}/\mathrm{SE}(\hat{\beta})$ having a standard normal distribution under the null hypothesis). These tests are equivalent to examining whether 95 percent confidence intervals around $\hat{\beta}$ include zero.

TABLE 3. Results of gene-environment interactions from the Yunnan Tin Corporation study (1992–1999), southern China, estimated by using case-control and case-only designs

Exposure	Case-control design*		Cases only†		Controls only‡	
	$\hat{I}$	95% confidence interval	$\hat{I}_{CO}$	95% confidence interval	$\hat{\psi}$	95% confidence interval
Tobacco use (median split)	0.443	0.170, 1.155	0.897	0.414, 1.941	2.026	1.148, 3.575
Alcohol consumption (yes/no)	0.370	0.143, 0.953	0.472	0.219, 1.018	1.276	0.733, 2.223

* Two main effects and interaction for full data.
† Association between tobacco use and *XRCC1* in cases.
‡ Association between tobacco use and *XRCC1* in controls.

We examine the robustness of the case-only design to various degrees of gene-environment (odds ratio) association. Specifically, we consider various nonzero log-odds ratios.

Estimation

Table 4 presents the bias and mean squared error for estimating the interaction effect, with parameters corresponding to estimates obtained from the tobacco-use analysis. For all log-odds ratios assessing independence, the estimated biases for the interaction effects estimated with the case-only design are similar to what would be expected from using equation 9. When the log-odds ratio assessing independence is zero, the case-only design is unbiased and very efficient. As anticipated by the analytical results described in the Efficiency of the Case-only Design section of this paper, there is an approximately 100 percent gain in efficiency for the case-only design over the case-control design, and the case-only design is as efficient as the cohort design. For a log-odds ratio assessing independence larger than 0.2 (odds ratio of only 1.22), estimates of the mean squared error (which penalizes for bias and variance) are larger for the case-only than the case-control design.

Testing

An important property of a statistical test is that it has the correct probability of a type I error (i.e., correct size of the test). This means that a test that rejects at the 0.05 significance level will be significant 5 percent of the time (over many simulations) when the null hypothesis is true. We used the case-only design to conduct a simulation (with 1,000 replications) examining the probability of a type I error for various levels of gene-environment association. For a $\log_e$-odds ratio assessing independence of 0, 0.2, and 0.5 (odds ratios = 1, 1.22, and 1.65), we estimated the probability of a type I error for a 0.05 significance level test as 0.041, 0.103, and 0.601, respectively. Thus, even with an odds ratio assessing independence of 1.22 (i.e., a very weak effect whose magnitude is smaller than the estimated odds ratio association of 1.28 for alcohol consumption and *XRCC1* presented in table 3), the type I error for testing interaction effects is inflated in the case-only design. Thus, the case-only design is very sensitive to the independence assumption. One either has to have a great deal of confidence in the independence assumption or evaluate it empirically with data on controls.

Various authors have proposed using data on controls to verify the independence assumption. We consider an adaptive procedure whereby we test for a gene-environment association by using the control data. If this test is not significant, then a case-only analysis is performed to test for an interaction. Otherwise, a case-control analysis is performed to test for the interaction. Table 5 shows the performance of this design with different gene-environment associations. We also present the power to detect different gene-environment associations by using data on controls. For varying degrees of gene-environment association, we estimated the probability of a type I error for detecting the gene-environment interaction with the sequential procedure. The estimated type I error probabilities are computed for both a 1-1 and 2-1 control-to-case ratio for a 0.05 significance level test of gene-environment association. The results demonstrate that the sequential procedure results in overinflated type I error rates. In addition, the power to detect a gene-environment interaction is low for $\log_e$-odds ratio associations of less than 0.5 (odds ratio = 1.65). Although the power of this test is increased with a higher significance level (e.g., a power of 0.44 to detect a $\log_e$-odds ratio of 0.2 with a 0.15 significance level), the type I error rate of the sequential procedure is still highly inflated.

The adaptive procedure, in which we first test for gene-environment association by using control data and then perform either a case-only analysis if the gene-environment

TABLE 4. Bias and MSE* for estimating log(multiplicative interaction) with design parameters estimated from the tobacco use analysis, by *XRCC1* interaction model from the Yunnan Tin Corporation study, southern China, 1992–1997†

	Design					
logψ	Cohort		Case-control		Case only	
	Bias	MSE	Bias	MSE	Bias	MSE
–0.3	–0.012	0.067	–0.035	0.121	–0.272	0.137
0	–0.015	0.056	–0.021	0.099	0.025	0.053
0.2	–0.012	0.058	–0.001	0.102	0.228	0.104
0.3	–0.005	0.052	–0.011	0.112	0.334	0.161
0.5	0.008	0.060	–0.008	0.112	0.544	0.353

* MSE, mean squared error, defined as MSE = Bias^2 + Variance; thus, the variance can be written as Variance = MSE – Bias^2.

† Data are simulated by using the model $\text{logit}P(D = 1) = -3.7 + 1.13E + 0.054G - 0.815GE$. The intercept term in the simulation model was chosen on the basis of a 3.7% lung cancer incidence in the study. The intercept was chosen to correspond to the estimate of incidence from the cohort study (339 cases out of 9,143 participants). All other parameters were chosen from the parameter estimates obtained by fitting a logistic model with environment-by-gene interactions to the case-control data.

TABLE 5. Performance of an adaptive procedure in which gene-environment association was tested for by using the control data*

$\log\psi$	1-1 control/case		2-1 control/case	
	Power for association test	Type I error for testing interaction	Power for association test	Type I error for testing interaction
–0.3	0.259	0.171	0.448	0.147
0	0.040	0.044	0.040	0.045
0.2	0.113	0.149	0.239	0.134
0.3	0.264	0.198	0.466	0.167
0.5	0.596	0.291	0.883	0.144

* We perform a case-only analysis if the test is nonsignificant at the 0.05 significance level and a case-control analysis otherwise. Design parameters were estimated from the tobacco-use-by-*XRCC1* interaction model fit to the Yunnan Tin Corporation study data. Data were simulated from the following model: $\text{logit}P(D = 1) = -3.5 + 0.682E - 0.383G$. Both the power to detect a gene-environment association in the controls and the type I error for testing the interaction at the 0.05 significance level are presented.

association is not significant or a case-control analysis if the association is significant, can result in a highly overinflated type I error. As an alternative, we examined the performance of an adaptive design in which the decision to perform a case-control or case-only analysis was based on the estimate of the gene-environment association odds ratio for controls. The performance of this design was evaluated for three cutoff values of $|\log_e\hat{\psi}| = 0.1, 0.2,$ and 0.3, when the true gene-environment associations were $\log_e\psi = 0.2$ or 0.5, and under a 1-1 control-to-case ratio. For a gene-environment association of $\log_e\psi = 0.5$, the type I error rate was 0.061, 0.082, and 0.143 for cutoff values of $\log_e\hat{\psi} = 0.1, 0.2,$ and 0.3, respectively. For a smaller gene-environment association of $\log_e\psi = 0.2$, the type I error rate was 0.092, 0.110, and 0.131 for the three cutoff values of 0.1, 0.2, and 0.3. Although this adaptive design is close to the nominal 0.05 level when the 0.1 threshold was used, the case-only analysis was chosen too rarely over the case-control analysis (3.5 percent and 21.6 percent for $\log_e\psi = 0.5$ and 0.2, respectively) to show an efficiency advantage of the adaptive design over case-control analysis.

DISCUSSION

Various authors have highlighted advantages of the case-only design in detecting gene-environment interactions in observational studies (1–6). Our work points out the need for caution, since inferences made with the case-only design can be highly sensitive to the often-unverified assumption of independence between the environmental exposure and the genetic marker. As expected, our simulations show that, when the assumptions are met, the case-only design results in more efficient estimates and more powerful tests than a case-control design with a similar number of cases. Unfortunately, they also reveal the sensitivity of the case-only design to even small amounts of gene-environment association.

We do not recommend discarding information from existing controls in a case-control study to use the more powerful case-only analysis. Relying on tests of independence in the controls is not effective because of their generally low power to detect meaningful departures from independence; our proposed adaptive procedures, which use a case-only or case-control analysis depending on the evidence of the independence assumption, resulted in inflated type I error probabilities even for small gene-environment associations.

The advantages of the case-only design are greater when there is a large quantity of empirical information about the gene-environment association or when selection of proper controls is difficult or impossible. Evidence of the gene-environment association could be used directly or in a sensitivity analysis to correct for a sizable gene-environment association by scaling the case-only estimate by the reciprocal of the gene-environment association (equation 3) or to support the independence assumption if the gene-environment odds ratio is near 1. For example, Marcus et al. (7) used the case-only design to test for a genotype by smoking interaction for the risk of bladder cancer. By using a large meta-analysis with approximately 2,000 controls, they demonstrated that this genotype is essentially independent from smoking. In such situations, the case-only design is not only more efficient, it is also less costly and uses fewer valuable resources than a traditional case-control study. Of course, the case-only design is subject to methodological problems such as uncontrolled confounding, exposure misclassification, and nonresponse bias. For example, for cultural reasons, only 30 percent of the cases in the YTC study were willing to provide blood for DNA typing (11). In its favor, the case-only design is immune to bias from poor control selection (13) and to exposure misclassification that is differential by disease status. However, the case-only design would be susceptible to bias if misclassification of exposure varied by mutational status.

The association between an environmental factor and a genetic marker in the controls could be due to genes associated with or causal for behaviors such as smoking and alcohol drinking. For example, polymorphisms of dopamine receptor genes have been shown to be associated with smoking habits (14). In addition, polymorphisms can be associated with behavior through linkage equilibrium (15). Furthermore, an association may be induced by ignoring an important stratification variable (i.e., uncontrolled confounding) that relates to both factors. For example, the environmental factor and the genetic marker may each be independent when subjects are young but dependent when they are older because of differential survival in the population if death is not rare (16). Alternatively, dependence could possibly be induced by differential participation in the study, even if the factors are independent in the population.

In this paper, we have illustrated the robustness of the case-only design to the independence assumption with two exposure variables (tobacco use and alcohol consumption) from the YTC study. These two exposures show the most extreme departure from independence among 10 studied. Thus, care must be taken in interpreting the empirical evidence for or against the independent assumption. For example, there is no known reason for an association between tobacco use and the *XRCC1* genotype. The result may be spurious (because of the multiplicity in screening 10 exposure variables) or possibly due to a linkage disequilibrium between the *XRCC1* Arg194Trp polymorphism and some other behavior-modifying genotype. In either case, our results demonstrate the potential problem of a case-only analysis when it is uncertain whether a gene-environment association exists.

The YTC study provides a good illustration of the limitations of the case-only design. Parameter estimates for the simulations were based on estimates from the YTC study. Thus, we should be cautious about generalizing the degree of bias and overinflation to other studies. Nonetheless, this limitation does not detract from our basic message that the case-only design may perform poorly when a gene-environment association exists. For simplicity, we ignored the matching (incidence density sampling) in the case-control analysis (table 3), and we selected controls from noncases at the end of the study in the simulations. Schmidt and Schaid (16) demonstrated that the case-only design may be biased when controls are selected either from survivors at the end of the study or by incidence density sampling. This bias is negligible in the YTC study (and in our simulations), where the risk of disease was low and the gene under study conferred only a moderately increased risk.

Our simulations suggest that the type I error is inflated (too many false positives) when there are departures from the independence assumption. Thus, one potential use of the case-only design is as an early screen for large gene-environment interactions in which all positive findings will be confirmed by using case-control or cohort studies. However, the case-only design may have low power to screen for gene-environment interactions when the associations of the environmental factor and the genetic trait with the gene-environment interaction are in the opposite direction (e.g., $\log_e I$ and $\log_e \psi$ in equation 9 have a different sign but a similar magnitude).

Sometimes the case-only or related analyses are the only practical option. In a case-control study of Jewish women in Israel, Modan et al. (10) investigated the important etiologic and public health question of whether use of oral contraceptives protects *BRCA1* and *BRCA2* carriers against ovarian cancer. Because of the small numbers of controls (13/751) who tested positive, even in a population in which carriers are relatively frequent, Modan et al. relied on an assumption of independence between carrier status and use of oral contraceptives rather than trying to estimate the effects of oral contraceptive use in a carrier stratum with only 13 controls. These authors noted that their analysis still provides important evidence about a question that is difficult to address in other settings.

In conclusion, the case-only design may be useful for assessing departures from multiplicative interaction when there is strong empirical evidence for the independence between the genetic marker and the environmental exposure, but estimates, tests, and confidence intervals should be interpreted cautiously in the absence of such evidence. Because the lack of robustness to the independence assumption outweighs the gain in power obtained with the case-only analysis, reliance on cases only for assessing multiplicative interactions is not a fully satisfactory substitute for a full case-control design and analysis. The case-only approach is nonetheless a useful tool, if used cautiously, for assessing interaction when the independence assumption is justified by empirical evidence or when selection of appropriate controls is difficult or impossible.

REFERENCES

1. Piegorsch WW, Weinberg CR, Taylor JA. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. Stat Med 1994; 13:153–62.
2. Khoury MJ, Flanders WD. Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: case-control studies with no controls! Am J Epidemiol 1996;144: 207–13.
3. Andrieu N, Goldstein AM. Epidemiologic and genetic approaches in the study of gene-environment interactions: an overview of available methods. Epidemiol Rev 1998,20:137–47.
4. Weinberg CR, Umbach DM. Choosing a retrospective design to assess joint genetic and environmental contributions to risk. Am J Epidemiol 2000;152:197–203.
5. Umbach DM, Weinberg CR. Designing and analyzing case-control studies to exploit independence of genotype and exposure. Stat Med 1997;16:1731–43.
6. Yang Q, Khoury MJ, Flanders WD. Sample size requirements in case-only designs to detect gene-environment interaction. Am J Epidemiol 1997;146:713–20.
7. Marcus PM, Hayes RB, Vineis P, et al. Cigarette smoking, *N*-acetyltransferase 2 acetylation status, and bladder cancer risk: a case-series meta-analysis of a gene-environment interaction. Cancer Epidemiol Biomarkers Prev 2000;9:461–7.
8. Bennett WP, Alavanja MCR, Blomeke B, et al. Environmental tobacco smoke, genetic susceptibility, and risk of lung cancer in never-smoking women. J Natl Cancer Inst 1999;91:2009–14.
9. Infante-Rivard C, Labuda D, Krajinovic M, et al. Risk of childhood leukemia associated with exposure to pesticides and with gene polymorphisms. Epidemiology 1999;10:481–7.
10. Modan B, Hartge P, Hirsh-Yechezkel G, et al. Parity, oral contraceptives, and the risk of ovarian cancer among carriers and noncarriers of a *BRCA1* or *BRCA2* mutation. N Engl J Med; 2001;345:235–40.
11. Ratnasinghe L, Yao SX, Tangrea JA, et al. Polymorphisms of the DNA repair gene *XRCC1* and lung cancer risk. Cancer Epidemiol Biomarkers Prev 2001;10:119–23.
12. Breslow NE, Day NE, eds. Statistical methods in cancer research. Vol 1. The analysis of case-control studies. Lyon, France: International Agency for Research on Cancer, 1980. (IARC scientific publication no. 32).
13. Wacholder S, McLaughlin JK, Silverman DT, et al. Selection of controls in case-control studies 1. Am J Epidemiol 1992;135:1019–50.
14. Spitz MR, Shi H, Yang F, et al. Case-control study of the D2 dopamine receptor gene and smoking status in lung cancer patients. J Natl Cancer Inst 1998;90:358–63.
15. Wu X, Hudmon KS, Detry MA, et al. D2 dopamine receptor gene polymorphisms among African-Americans and Mexican-Americans: a lung cancer case-control study. Cancer Epidemiol Biomarkers Prev 2000;9:1021–6.
16. Schmidt S, Schaid DJ. Potential misinterpretation of the case-only study to assess gene-environment interaction. Am J Epidemiol 1999;150:878–85.